

# Chemistry42: An AI-Driven Platform for Molecular Design and Optimization

Yan A. Ivanenkov, Daniil Polykovskiy, Dmitry Bezrukov, Bogdan Zagribelnyy, Vladimir Aladinskiy, Petrina Kama, Alex Aliper, Feng Ren, and Alex Zhavoronkov\*



Cite This: *J. Chem. Inf. Model.* 2023, 63, 695–701



Read Online

ACCESS |



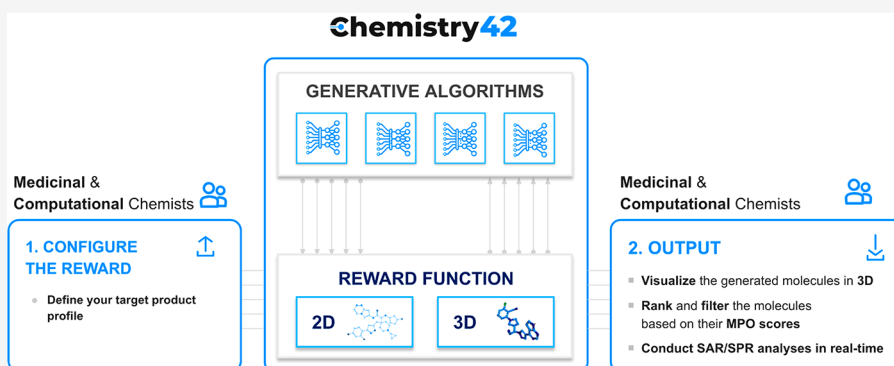
Metrics & More



Article Recommendations



Supporting Information



**ABSTRACT:** Chemistry42 is a software platform for *de novo* small molecule design and optimization that integrates Artificial Intelligence (AI) techniques with computational and medicinal chemistry methodologies. Chemistry42 efficiently generates novel molecular structures with optimized properties validated in both *in vitro* and *in vivo* studies and is available through licensing or collaboration. Chemistry42 is the core component of Insilico Medicine's **Pharma.ai** drug discovery suite. Pharma.ai also includes PandaOmics for target discovery and multiomics data analysis, and inClinico—a data-driven multimodal forecast of a clinical trial's probability of success (PoS). In this paper, we demonstrate how the platform can be used to efficiently find novel molecular structures against DDR1 and CDK20.

## INTRODUCTION

Deep Learning (DL) has proven to be very effective in speech and image recognition. This is because DL-based architectures are uniquely suited for the automatic identification of patterns within complex, nonlinear data sets without the need for manual feature engineering. DL methods have recently been adapted to successfully overcome limitations inherent in the standard techniques used for small molecule design.<sup>1–3</sup> These adaptations offer exciting possibilities for the development of new methods that efficiently explore uncharted chemical space.

Insilico Medicine was one of the first groups to publish a method that uses a deep adversarial model for new compound generation.<sup>4</sup> Since then, DL-based architectures that combine generative algorithms with reinforcement learning (RL) have been developed and applied in chemistry and pharmacology to generate novel molecular structures with predefined properties.<sup>2</sup> Especially encouraging is the recent progress in the *de novo* design of active molecules that have been validated in both *in vitro* and *in vivo* assays.<sup>5,6</sup> Generative chemistry is now one of the fastest-growing areas in drug discovery.<sup>2,7,8</sup> The Chemistry42 platform has been routinely and successfully used at Insilico Medicine to drive the drug discovery process in several

therapeutic areas (<https://insilico.com/pipeline>) and has evolved significantly during the past years.<sup>9</sup> In the following sections, we describe the key features of the Chemistry42 platform.

## OVERVIEW OF THE GENERATIVE CAPABILITIES OF THE CHEMISTRY42 PLATFORM

Chemistry42 is a platform that connects state-of-the-art generative AI algorithms with medicinal and computational chemistry expertise and best engineering practices. It was launched in 2020 and has been used by over 20 pharmaceutical companies, over 15 external programs, and over 30 internal programs.

The main objective of this platform is to accelerate the design of novel molecules with user-defined properties. The general

**Received:** September 22, 2022

**Published:** February 2, 2023

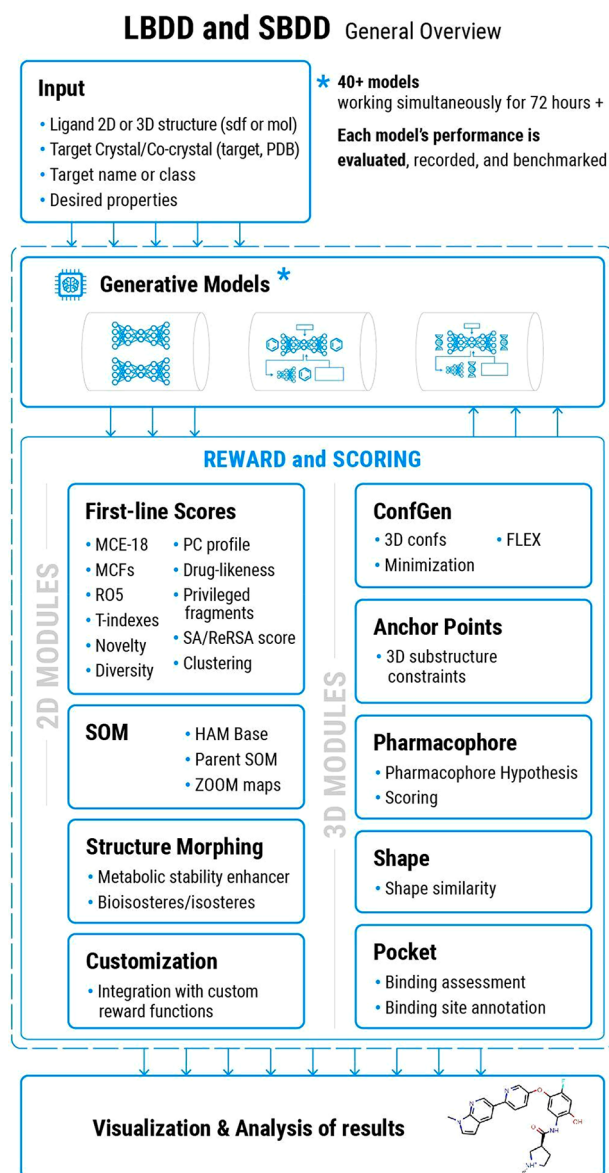


workflow for Chemistry42 is illustrated and described in Figure 1.

Generative experiments are created using the user-friendly web-based interface of Chemistry42 and can be started using ligand- or structure-based drug design workflows depending on the available information for the target of choice. The Ligand-Based Drug Design (LBDD) approach requires a 2D or 3D ligand structure as input in an .sdf file, a SMILES string, or the molecule can be sketched directly on the platform using the handy sketcher panel. A pharmacophore hypothesis can also be added as needed and created manually using a widget or automatically within the platform. In the Structure-Based Drug Design (SBDD) approach, the structure of a protein target, either in the *apo* format or in complex with a ligand, must be uploaded to the platform as a prepared .pdb file. One can pick either the pocket around the ligand (ligand binding site) or select one from the set of alternative pockets indicated by the *Pocket Scanner Module*. As with the case of LBDD, a pharmacophore hypothesis can also be added as needed (Figure 2). To complete the configuration of a generation experiment, the user defines acceptable ranges for multiple properties (e.g., physicochemical properties and diversity) of the generated structures. The user can prioritize reward modules by adjusting their weights and specify how restrictive the modules should be by adjusting corresponding thresholds. In both LBDD and SBDD approaches, advanced options enable the user to specify and fine-tune reward modules and which generative models should be used in an experiment. Hit expansion, hit-optimization, and Fragment-Based Drug Design (FBDD) workflows are also available on the platform through functionality called *Anchor points*. With *Anchor points*, users can fix in 3D space specified cores or R-groups of a hit-molecule while the rest of the molecule varies during a generative experiment. *Anchor points* also support multiple reference substructures by editing atom types to include alternatives (supported by SMARTS). For example, the user can specify whether they would like to see nitrogen or carbon in an aromatic ring. Autoconfiguration is a quick way of adjusting all the parameters automatically, based on the provided input data. To see how the platform can be autoconfigured based on the input ligand structure and properties see S1 section in the SI.

The generative pipeline in Chemistry42 engages an asynchronous ensemble of proprietary generative models. These carefully curated and selected algorithms have diverse architectures that implement distinct strategies. The platform takes advantage of multiple machine learning models and molecular representations for different generative scenarios to maximize each model's contribution and the platform's efficiency. For example, some models focus on the exploration of the chemical space, while being tailored to improve these explored structures. In the current version of Chemistry42, there are over 40 generative models, including generative autoencoders,<sup>5,10</sup> generative adversarial networks,<sup>4,11,12</sup> flow-based approaches,<sup>13</sup> evolutionary algorithms,<sup>14</sup> language models,<sup>15</sup> and others. These models employ different molecular representations—string-based, graph-based, and 3D-based.

It is essential to understand and stimulate the interplay of multiple generative models. As such, rather than treating these algorithms as black-box solutions, we provide deep domain-specific analytics to understand the advantages and drawbacks of each approach. Combining various state-of-the-art machine learning methods, Chemistry42 delivers diverse, high-quality molecular structures within hours. As the structures are



**Figure 1.** A schematic representation of the three-step workflow for a *de novo* generative experiment using the Chemistry42 platform. In the first step, on a secure and company-specific instance of the software, users upload their data and configure the platform with the desired properties for the generated structures. The second step involves running the platform where an ensemble of 40+ generative models functions in parallel to generate the novel structures—this step is called the generation phase. A variety of filters scrutinize the generated molecular structures in the generation phase. The molecular structures are then subjected to multiple sets of reward and scoring modules, classified as either 2D or 3D modules, that dynamically assess the generated structures' properties according to the predefined criteria. Additional custom scoring modules (such as ADME predictors) can also be integrated into the reward pipeline to prioritize the generated structures. These modules form the backbone of Chemistry42's multiagent reinforcement learning (RL)-based generation protocol. Generated structures' scores are fed back to the generative models to reinforce them and guide the generative process toward high-scoring structures—this is called the learning phase. The final step is analysis. The generated structures are automatically ranked according to customizable metrics based on their predicted properties, including synthetic accessibility, novelty, diversity, etc. The platform also provides users with interactive tools to monitor generative model performance.

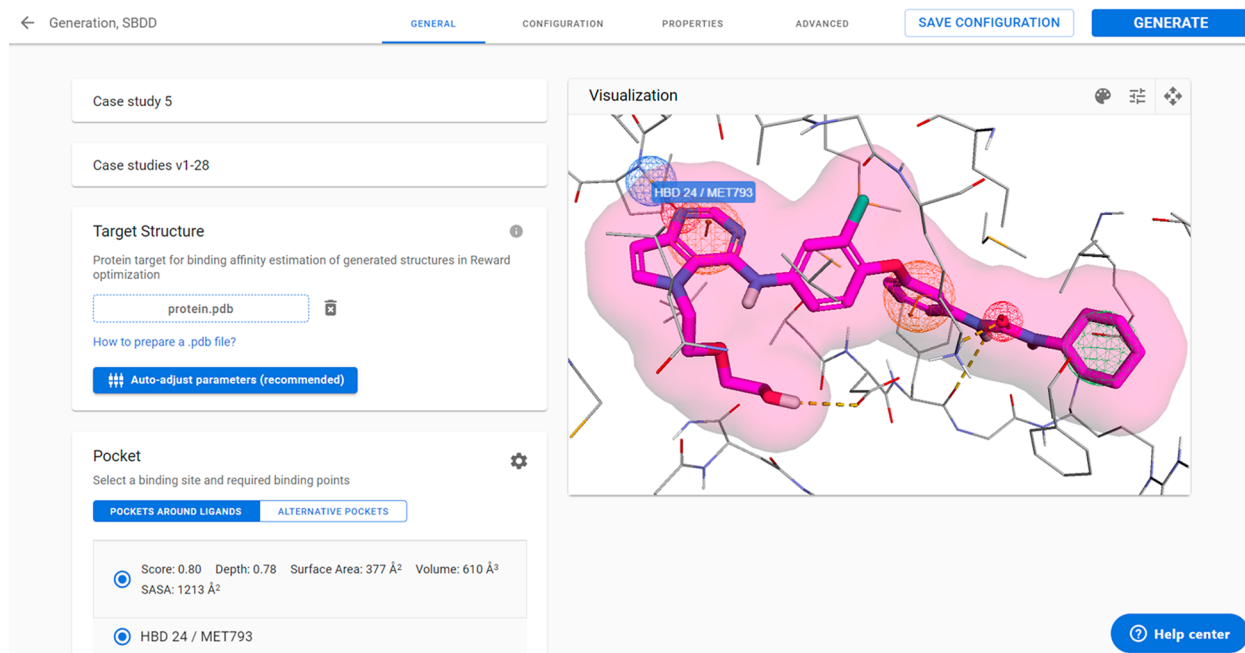


Figure 2. Chemistry42 interface for configuring an SBDD generative experiment.

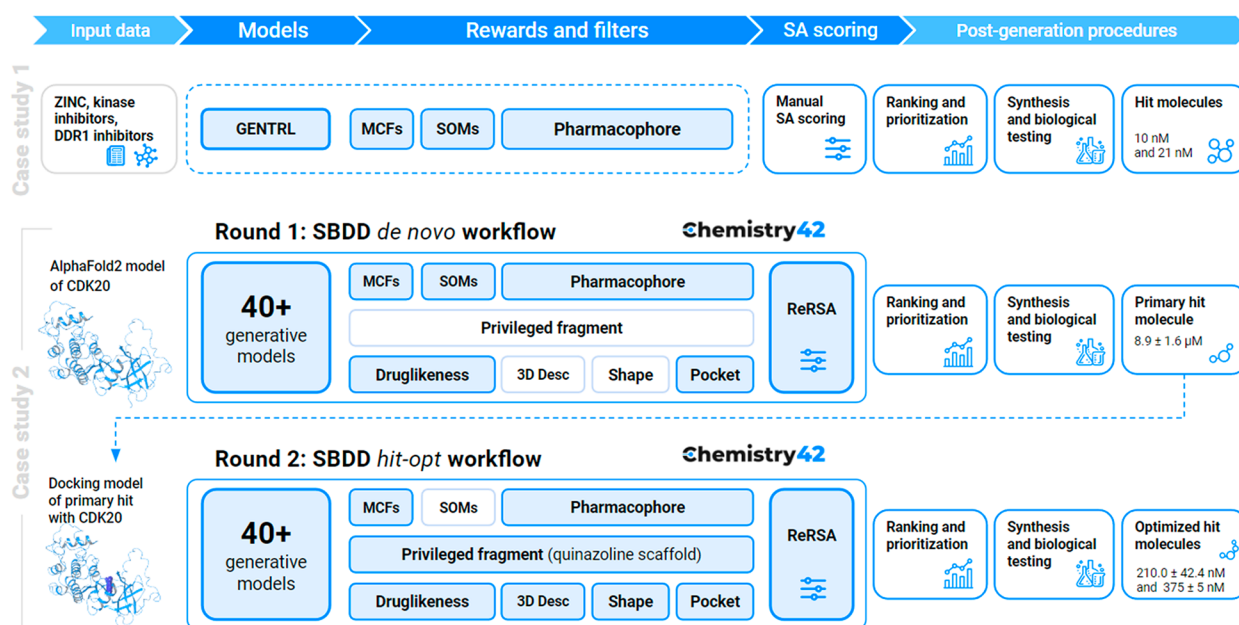
generated, they are dynamically assessed using the reward and scoring modules in the platform.

The reward and scoring modules used in Chemistry42 for RL-based generations are either two-dimensional (2D) or three-dimensional (3D) (Figure 1). The 2D modules are composed of multiple scores and in-house *Medicinal Chemistry Filters* (MCFs) that are used to assess the generated structures. In the current version of Chemistry42, the MCFs include a set of over 460 in-house structure-based rules that exclude “bad” structures, i.e., those that contain structure alerts, PAINS<sup>16</sup> or functional groups that are reactive, unstable or potentially toxic. The *Medicinal Chemistry Evolution* (MCE-18) function is a unique molecular descriptor that scores structures by novelty in terms of their cumulative  $sp^3$  complexity.<sup>17</sup> Other 2D modules include Lipinski’s Rule of Five (Ro5),<sup>18</sup> *Drug-likeness* and *Weighted atom-type portion descriptors*, a rule-based filter that constitutes a set of rules to eliminate structures with an unbalanced number of heteroatoms and aromatic atoms. *Novelty* scores assess the 2D similarity between the generated structures and the reference data set (that can be customized) to improve novelty. *Drug-likeness* is estimated using a set of extended rules. The synthetic accessibility (SA) of the generated structures is assessed using the *Retrosynthesis Related Synthetic Accessibility* (ReRSA) score.<sup>19</sup> ReRSA is an improved fragment-based SA estimation method that is based on the fragmentation of the generated structure from a retrosynthesis perspective, which results in a more accurate estimation of SA. ReRSA also takes into account the space of commercially available building blocks and rewards a structure if it can be converted into existing building blocks (BBs). By default, ReRSA exploits the approximately 200,000 commonly used BBs. *Diversity* assessments and clustering metrics are performed for the generated structures using a custom similarity function. Tracking the *Diversity* of the generated structures provides a means of understanding how structurally diverse the generated molecules are based on the number of generated chemotypes following clustering. *Privileged Fragments* (PFs) are defined structural motifs that contribute to the activity of a target or target class.<sup>20</sup> PFs functionality is most

useful in two types of generative design workflows. The first involves defining 2D PF substructure(s) that will be found in all generated structures with no predefined positioning in 3D space. This is useful if you only have an apo-protein structure with no reported inhibitors. For example, if your target is the apo structures of the novel kinase, 2D PFs of hinge binders can be used in the generative experiments to navigate the generation into well-established chemical space. The second workflow involves the use of *Anchor Points*—essentially 3D privileged fragments. Here, the presence of the substructure of interest is essential in either a protein–ligand complex (SBDD mode) or a 3D conformation of a ligand (LBDD mode). The self-organizing maps (SOM) *Classifier Module* (general SOM map  $100 \times 100$ ) is used to drive the generation of molecular structures toward the chemical space corresponding to a specified target class. Since the general SOM contains neurons with the classification power below a predefined threshold for a selected category of molecules, all the reference molecules from such neurons are collected and then subjected to automatically generated ZOOM maps of an adapted size to achieve reliable classification accuracy. The data set used to train the SOMs *Classifier Module*,<sup>21</sup> and ZOOM maps are called the *Hierarchical Active Molecules* (HAM) data set. The HAM data set consists of data from 800k+ experimentally validated molecules with IC50s of 10  $\mu$ M or less. The Structure Morphing module contains two components: a rule-based *Metabolic Stability Enhancer*<sup>22</sup> that addresses metabolic instability caused by potential sites of metabolism in the generated structures and the *Bioisostere Module* that performs bioisosteric/isosteric transformations.<sup>23</sup>

Following the assessment of the generated structures with the 2D modules, multiple 3D modules are deployed for further assessment. The *ConfGen Module* is the first 3D module. It produces a conformational ensemble for each generated structure. The *ConfGen Module* generates conformational ensembles through an in-house set of rules and predefined substructure geometries based on small molecule cocrystal X-ray data followed by energy minimization using Insilico’s proprietary force field. A flexibility assessment (*FLEX score*) is used to





**Figure 3.** DDR1 inhibitors generation was performed in 2018 by the GENTRL model in Case study 1 (above). The supporting postgeneration modules (MCFs, SOMs, and pharmacophores) were utilized to narrow down the generation output. Chemistry42 is an integrated platform released in 2020, where more than 40 generative models work together and get information from the reward modules and filters to produce molecular structures with desired properties. The newer version of the platform enables the exploitation of .pdb files of protein–ligand complexes and apo-structures as input data. The CDK20 model from AlfaFold2 was used in two sequential rounds of generation. The first round was focused on the *de novo* design of potential CDK20 inhibitors from apo-structure, while the second round exploited a hit-opt workflow starting from the primary hit identified in the first round.

rank molecular structures by intrinsic rigidity. Once the conformational ensembles have been generated, the *3D-Descriptors Module* evaluates the 3D similarity between the generated structures and a reference molecule (input ligand) using a set of calculated 3D-descriptors. The *Pharmacophore Module* then assesses if any of the generated conformations match the specified pharmacophore hypothesis, including all important binding points, distances, angles, and tolerance. If the *Anchor Points* module is used in the generation, it checks if the user-defined 3D substructures are present in the generated structure in the correct position and conformation. The *Shape Similarity Module* evaluates the 3D-shape similarity to a reference molecule using weighted Gaussian functions.<sup>24</sup> The final module focuses on positioning and scoring the generated structures to assess how well they fit the selected binding site (*Pocket Module*) and approximates the binding affinity with a Pocket-Ligand Interaction (*PLI*) score. The *PLI* score was trained on the PDBBind Refined Set v2020 (both  $K_i$  and  $K_d$  data were used).<sup>25</sup> The score takes into account hydrogen bonds,  $\pi$ -stacking,  $\pi$ -cation, XH- $\pi$ , hydrophobic interactions, as well as salt bridges and chelating bonds. The units of the *PLI* score are kcal/mol, where the more negative the value, the better the score. The list of most important scores and rewards for each individual workflow (e.g., *de novo* SBDD or FBDD) is available in the S2 section of SI. The tabulated list of all mentioned modules, scores, and rewards, as well as default ranges, values, and options, for them can be found in the S3 section of SI.

The user can specify how long they want to run the generative experiment. In most cases, we observe convergence after 72 h. During a generative experiment, the performance of each generative model is monitored and recorded. This allows the user to follow the progress of their experiments in real-time from start to completion. An example of the generation monitoring at the early stage and after the completion can be found in S4

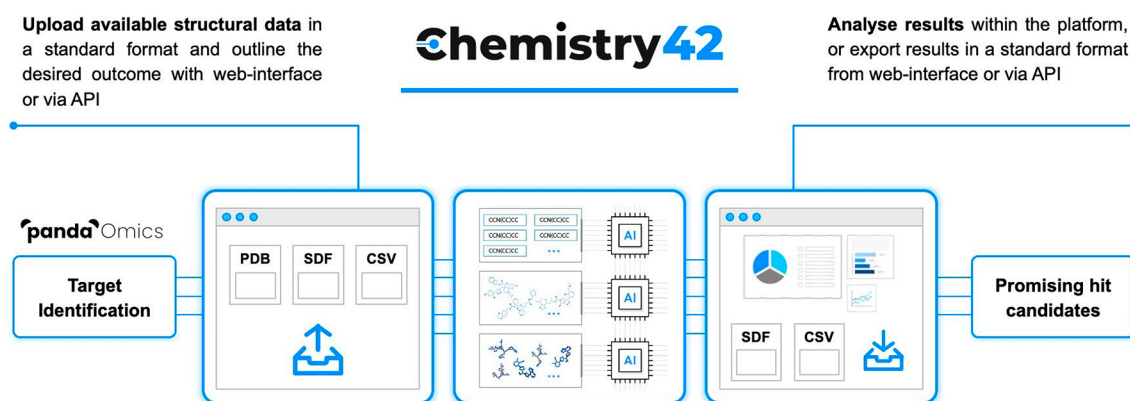
section of SI. The generated structures are automatically evaluated and ranked according to metrics incorporated in the modules that are integrated into the platform. All corresponding data, including scores, molecular structures, and generative model performance, are stored and accessible on the results page of the platform. Once a generative experiment is complete, the results can be analyzed through an interactive interface.

A mediocre user can get valuable results (1–5 novel molecules for synthesis) even from the first run of the generation for both SBDD, LBDD, and other various workflows. The obtained results and their subsequent analysis can help to configure the second run in a more specific way to get more expected results. Usually, at the second run the user can utilize some ideas that have been provided by the platform at the first run (e.g., add some privileged structures at the second run). An advanced user who is deeply familiar with the functionality and outcomes of the platform can configure the platform and get expected results (10–20 of novel molecules for synthesis) even in the first run for a new project.

A user is provided at the platform with an online well-documented manual describing sample virtual case studies for different workflows mentioned in SI section S2, step-by-step instructions for these case studies and sample outcomes as .sdf files and their analysis by medicinal and computational chemists. These case studies can help a fresher to accept the more appropriate strategy to exploit the platform for their own project purposes.

## ■ BENCHMARKING THE GENERATIVE MODELS IN CHEMISTRY42

The performance of all generative models used in an experiment is monitored and recorded by a benchmarking system based on the Molecular Sets (MOSES) system. MOSES assesses the



**Figure 4.** PandaOmics and Chemistry42 platforms integrated into your drug discovery pipeline. The interoperability of these platforms allows an efficient interaction between target identification and de novo small molecule generation.

performance of each model and the reward components, including novelty, diversity, and others, during the generation and once the experiment is completed.<sup>26</sup> Based on the provided analytics, users can analyze each model's performance. A record of the results and training data is kept during the experiment and stored to ensure that reproducibility and monitoring are both simple and feasible. An example of MOSES-based analytics for a generation is available in the S5 of [Supporting Information](#).

## CASE STUDIES

Early versions of the generative pipeline (prior to the launch of Chemistry42), were used to demonstrate the ability of generative algorithms to design experimentally validated, druglike molecular structures (see [Figure 3](#)).

**Case Study 1: GENTRL Model Enabled the Rapid Generation of Potent DDR1 Inhibitors.** The GENTRL model and postgeneration protocols are the ancestors of the current architecture of the Chemistry42 platform. In the seminal and widely discussed study<sup>5</sup> the model developed in 2018 generated experimentally validated potent DDR1 kinase inhibitors. GENTRL was trained on the ZINC data set and then fine-tuned on reported DDR1 inhibitors and a publicly available kinase inhibitor data set. The obtained structures were then passed through structural filters to eliminate structures with reactive groups, PAINS, and other alerts. This functionality was further converted into the MCFs module of the Chemistry42 platform. The number of structures was further reduced by clustering and the selection of the most diverse members of each cluster. This smaller subset of structures was then assessed on kinase SOMs and by pharmacophore hypotheses, constructed from reported cocrystals of DDR1 with its inhibitors. The remaining structures were subject to a random selection of 40 that were further subjected to a manual synthetic accessibility evaluation. Of the 40 structures, six were nominated for synthesis and biological assessment. By day 35, the compounds had been successfully synthesized and tested *in vitro* for the inhibition of DDR1 enzymatic kinase assay. More than half of the compounds were found to be active ( $IC_{50} < 1 \mu M$ ), including two two-digit nanomolar inhibitors (10 nM and 21 nM), while two compounds showed no activity in the assay.

**Case Study 2: The Use of AlphaFold2 Generated 3D Protein Structures to Generate Hit-Molecules.** We have recently demonstrated the use of an AlphaFold2<sup>27</sup> predicted protein structure in an SBDD case study of Cyclin-Dependent Kinase 20 (CDK20) inhibitors.<sup>28</sup> The selected target, CDK20

(also known as cell cycle-related kinase, CCRK), was identified as a promising drug target for hepatocellular carcinoma by the PandaOmics<sup>29</sup> software. The absence of structural information for both protein and reported tool compounds makes CDK20 an ideal candidate for the validation of our AlphaFold2+Chemistry42 approach. The CDK20 AlphaFold2 model (AF-Q8IZL9-F1-model\_v1) was used as input for Chemistry42 in the SBDD mode to produce new molecular structures that would inhibit CDK20. The *Pocket Scanner Module* mapped the ATP-binding site as an ideal pocket of choice to generate potential inhibitors. To navigate the generation into the chemical space common for kinases and specifically CDK inhibitors a classic hinge binder pharmacophore and a SOM trained on the known CDK inhibitors were used. The *Novelty* filter was engaged to ensure that generated molecules will not share similar structures to the existing molecules from the CDK subset of the *HAM data set*. Also *Shape* and *3D-descriptors* modules were disabled at the first de novo stage, since these modules are not applicable in the absence of the template ligand, when the apo-structure of modeled protein is used (see disabled modules at the [Figure 3](#)). In total, 8918 molecules were designed by the generative pipeline. After molecular docking and visual pose inspection, 54 molecules with diverse hinge binder scaffolds were prioritized and seven compounds were nominated for synthesis based on their scores. Among these compounds, a primary hit containing a quinazoline ring with  $K_d$  value of  $8.9 \pm 1.6 \mu M$  in a CDK20 binding assay was discovered. The experimental details of the assay are available in the section S6 of [SI](#). Hit-optimization at the second stage using the 2D PFs functionality was deployed to maintain the quinazoline scaffold and explore the R-group space to improve the binding affinity of the identified hit-molecule. From this hit-optimization generation that exploited the same pharmacophore from the first round and did not use SOM scoring, six molecules were short-listed for synthesis based on their scores, and two compounds showed remarkable improvement in potency resulting in  $K_d$  values of  $210.0 \pm 42.4$  nM and  $375 \pm 5$  nM, respectively. The default set of MCFs and default ranges for *Drug-likeness* related properties (see section S3 in [SI](#)) were applied for both generations. Altogether this work demonstrated the synergy between AI methods supporting target identification (PandaOmics), protein folding (AlphaFold2), and generative chemistry (Chemistry42) in their ability to effectively contribute to a digital drug development process when structural data are limited. The structures of all 7 compounds selected from the first round, as well as those of

optimized hit-compound and their similarity analysis using ChEMBL tools, are available in the section S7 of SI.

## CHEMISTRY42 INTEROPERABILITY

Chemistry42 is accessible through a web interface built on top of a distributed cloud platform with scalable cloud architecture. The implementation integrates a variety of features aimed at optimizing performance, including cluster management with Kubernetes, multiple flexible workflows, and integrated monitoring and logging. The structure and interoperability of the Chemistry42 platform allow its deployment on either the AWS or Azure cloud or as a SaaS solution ([chemistry42.com](https://chemistry42.com)). For either deployment scenario, the platform can be integrated into already established workflows.

Chemistry42 can be connected to Insilico Medicine's bioinformatics web service PandaOmics (<https://pandaomics.com>) (Figure 4). PandaOmics is a comprehensive computational suite for the analysis of -omics data that provides access to information ranging from disease signatures to prospective targets and existing drugs. PandaOmics combines classic bioinformatics methods with signaling pathway analysis using the iPANDA algorithm.<sup>29–33</sup> PandaOmics also provides access to an AI-powered toolkit including deep feature selection for pathway reconstruction, a pathway scoring engine, causal inference, a deep-learned transcriptional response scoring engine, and an activation-based scoring engine. This multimodal approach combines big data, chemistry, biology, and medicine and allows a complete characterization of the interplay between molecular structures, properties, alteration in biological samples, and drug response required for target discovery.

## CONCLUSION

The Chemistry42 platform (<https://chemistry42.com>) is a customizable working environment that offers state-of-the-art AI technologies developed for *de novo* molecular design. The flexible, user-friendly interface makes Chemistry42 accessible to medicinal and computational chemists, AI experts, and other scientists working in the field of drug discovery. The collaborative nature of Chemistry42 enables and fosters relationships between different scientific communities and facilitates the decision-making process—a process which is exceptionally demanding in the field of drug design.

## ASSOCIATED CONTENT

### Data Availability Statement

The Chemistry42 platform is commercially available to the public (<https://chemistry42.com>). Parts of the platform, such as the GENTRL algorithm, are available online <https://github.com/insilicomedicine/GENTRL>. Data for training the models is constructed from publicly available sources such as ChEMBL (<https://www.ebi.ac.uk/chembl/>).

### Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.jcim.2c01191>.

Example of autoconfiguration, the list of the most important scores and rewards for different workflows, the list of reward/scoring functions, visualization of generative models' performance during the generation, CDK20 Human CMGC kinase binding assay protocol, and similarity analysis using ChEMBL tools (PDF)

## AUTHOR INFORMATION

### Corresponding Author

Alex Zhavoronkov — Insilico Medicine Kong Kong Ltd., Pak Shek Kok, Hong Kong; [orcid.org/0000-0001-7067-8966](https://orcid.org/0000-0001-7067-8966); Email: [alex@insilico.com](mailto:alex@insilico.com)

### Authors

Yan A. Ivanenkov — Insilico Medicine Kong Kong Ltd., Pak Shek Kok, Hong Kong; [orcid.org/0000-0002-8968-0879](https://orcid.org/0000-0002-8968-0879)  
Daniil Polykovskiy — Insilico Medicine Canada Inc., Montreal, Quebec H3B 4W8, Canada; [orcid.org/0000-0002-0899-8368](https://orcid.org/0000-0002-0899-8368)  
Dmitry Bezrukov — Insilico Medicine Kong Kong Ltd., Pak Shek Kok, Hong Kong  
Bogdan Zagribelnyy — Insilico Medicine AI Limited, Abu Dhabi, UAE  
Vladimir Aladinskiy — Insilico Medicine AI Limited, Abu Dhabi, UAE  
Petrina Kamya — Insilico Medicine Canada Inc., Montreal, Quebec H3B 4W8, Canada  
Alex Aliper — Insilico Medicine AI Limited, Abu Dhabi, UAE  
Feng Ren — Insilico Medicine Shanghai Ltd., Shanghai 201203, China; [orcid.org/0000-0001-9157-9182](https://orcid.org/0000-0001-9157-9182)

Complete contact information is available at:

<https://pubs.acs.org/doi/10.1021/acs.jcim.2c01191>

### Notes

The authors declare the following competing financial interest(s): Y.A.I., D.P., D.B., B.Z., V.A., P.K., A.A., F.R., and A.Z. work for Insilico Medicine, a commercial artificial intelligence company that developed the Chemistry42 platform.

## ACKNOWLEDGMENTS

The authors gratefully acknowledge the valuable comments and suggestions made by Dr. Jiye Shi from UCB Pharma (Slough, UK).

## REFERENCES

- (1) Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The Rise of Deep Learning in Drug Discovery. *Drug Discovery Today* **2018**, *23* (6), 1241–1250.
- (2) Vanhaelen, Q.; Lin, Y.-C.; Zhavoronkov, A. The Advent of Generative Chemistry. *ACS Med. Chem. Lett.* **2020**, *11*, 1496.
- (3) Yang, X.; Wang, Y.; Byrne, R.; Schneider, G.; Yang, S. Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery. *Chem. Rev.* **2019**, *119* (18), 10520–10594.
- (4) Kadurin, A.; Aliper, A.; Kazennov, A.; Mamoshina, P.; Vanhaelen, Q.; Khrabrov, K.; Zhavoronkov, A. The Cornucopia of Meaningful Leads: Applying Deep Adversarial Autoencoders for New Molecule Development in Oncology. *Oncotarget* **2017**, *8* (7), 10883–10890.
- (5) Zhavoronkov, A.; Ivanenkov, Y. A.; Aliper, A.; Veselov, M. S.; Aladinskiy, V. A.; Aladinskaya, A. V.; Terentiev, V. A.; Polykovskiy, D. A.; Kuznetsov, M. D.; Asadulaev, A.; Volkov, Y.; Zholus, A.; Shayakhmetov, R. R.; Zhebrak, A.; Minaeva, L. I.; Zagribelnyy, B. A.; Lee, L. H.; Soll, R.; Madge, D.; Xing, L.; Guo, T.; Aspuru-Guzik, A. Deep Learning Enables Rapid Identification of Potent DDR1 Kinase Inhibitors. *Nat. Biotechnol.* **2019**, *37* (9), 1038–1040.
- (6) Chen, H.; Engkvist, O. Has Drug Design Augmented by Artificial Intelligence Become a Reality? *Trends Pharmacol. Sci.* **2019**, *40* (11), 806–809.
- (7) Schneider, G. Generative Models for Artificially-Intelligent Molecular Design. *Mol. Inform.* **2018**, *37* (1–2), 1880131.
- (8) Merk, D.; Friedrich, L.; Grisoni, F.; Schneider, G. De Novo Design of Bioactive Small Molecules by Artificial Intelligence. *Mol. Inform.* **2018**, *37* (1–2), 1700153.



- (9) Zhavoronkov, A. From Paper to Industrial-scale Platform: a 3-Year Behind the Paper Journey from GENTRL to Chemistry42. *Bioengineering*, Springer Nature. <http://bioengineeringcommunity.nature.com/posts/from-paper-to-industrial-scale-platform-a-3-year-behind-the-paper-journey-from-gentrl-to-chemistry42> (accessed 2022-09-21).
- (10) Polykovskiy, D.; Zhebrak, A.; Vetrov, D.; Ivanenkov, Y.; Aladinskiy, V.; Mamoshina, P.; Bozdaganyan, M.; Aliper, A.; Zhavoronkov, A.; Kadurin, A. Entangled Conditional Adversarial Autoencoder for de Novo Drug Discovery. *Mol. Pharmaceutics* **2018**, *15* (10), 4398–4405.
- (11) Kadurin, A.; Nikolenko, S.; Khrabrov, K.; Aliper, A.; Zhavoronkov, A. druGAN: An Advanced Generative Adversarial Autoencoder Model for de Novo Generation of New Molecules with Desired Molecular Properties in Silico. *Mol. Pharmaceutics* **2017**, *14* (9), 3098–3104.
- (12) Putin, E.; Asadulaev, A.; Ivanenkov, Y.; Aladinskiy, V.; Sanchez-Lengeling, B.; Aspuru-Guzik, A.; Zhavoronkov, A. Reinforced Adversarial Neural Computer for de Novo Molecular Design. *J. Chem. Inf. Model.* **2018**, *58* (6), 1194–1204.
- (13) Kuznetsov, M.; Polykovskiy, D. MolGrow: A Graph Normalizing Flow for Hierarchical Molecular Generation. *AAAI* **2021**, *35* (9), 8226–8234.
- (14) Devi, R. V.; Sathya, S. S.; Coumar, M. S. Evolutionary Algorithms for de Novo Drug Design – A Survey. *Appl. Soft Comput.* **2015**, *27*, 543–552.
- (15) Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.* **2018**, *4* (1), 120–131.
- (16) Baell, J. B.; Holloway, G. A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J. Med. Chem.* **2010**, *53* (7), 2719–2740.
- (17) Ivanenkov, Y. A.; Zagribelnyy, B. A.; Aladinskiy, V. A. Are We Opening the Door to a New Era of Medicinal Chemistry or Being Collapsed to a Chemical Singularity? *J. Med. Chem.* **2019**, *62* (22), 10026–10043.
- (18) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **2001**, *46* (1–3), 3–26.
- (19) Zagribelnyy, B.; Putin, E. O.; Fedorchenko, S. A.; Ivanenkov, Y. A.; Zhavoronkov, A. Retrosynthesis-Related Synthetic Accessibility. Patent 2021229454:A1, November 18, 2021. <https://patentimages.storage.googleapis.com/cc/f9/aa/90acd239be7c4c/WO2021229454A1.pdf> (accessed 2022-08-11).
- (20) Yet, L. *Privileged Structures in Drug Discovery: Medicinal Chemistry and Synthesis*; Methods and Principles in Medicinal Chemistry; John Wiley & Sons, 2018.
- (21) Kohonen, T. *Self-Organizing Maps*, 3rd. extended ed.; Springer Series in Information Sciences; Springer-Verlag, Berlin, Germany, 2001; Vol. 30.
- (22) Kirchmair, J.; Göller, A. H.; Lang, D.; Kunze, J.; Testa, B.; Wilson, I. D.; Glen, R. C.; Schneider, G. Predicting Drug Metabolism: Experiment And/or Computation? *Nat. Rev. Drug Discovery* **2015**, *14* (6), 387–404.
- (23) *Bioisosteres in Medicinal Chemistry*; Brown, N., Ed.; Methods and Principles in Medicinal Chemistry; John Wiley & Sons, 2012; Vol. 54.
- (24) Yan, X.; Li, J.; Liu, Z.; Zheng, M.; Ge, H.; Xu, J. Enhancing Molecular Shape Comparison by Weighted Gaussian Functions. *J. Chem. Inf. Model.* **2013**, *53* (8), 1967–1978.
- (25) Liu, Z.; Li, Y.; Han, L.; Li, J.; Liu, J.; Zhao, Z.; Nie, W.; Liu, Y.; Wang, R. PDB-Wide Collection of Binding Data: Current Status of the PDBbind Database. *Bioinformatics* **2015**, *31* (3), 405–412.
- (26) Polykovskiy, D.; Zhebrak, A.; Sanchez-Lengeling, B.; Golovanov, S.; Tatanov, O.; Belyaev, S.; Kurbanov, R.; Artamonov, A.; Aladinskiy, V.; Veselov, M.; Kadurin, A.; Johansson, S.; Chen, H.; Nikolenko, S.; Aspuru-Guzik, A.; Zhavoronkov, A. Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *Front. Pharmacol.* **2020**, *11*, 1931.
- (27) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Zidek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589.
- (28) Ren, F.; Ding, X.; Zheng, M.; Korzinkin, M.; Cai, X.; Zhu, W.; Mantsyzov, A.; Aliper, A.; Aladinskiy, V.; Cao, Z.; Kong, S.; Long, X.; Man Liu, B. H.; Liu, Y.; Naumov, V.; Shneyderman, A.; Ozerov, I. V.; Wang, J.; Pun, F. W.; Polykovskiy, D. A.; Sun, C.; Levitt, M.; Aspuru-Guzik, A.; Zhavoronkov, A. AlphaFold Accelerates Artificial Intelligence Powered Drug Discovery: Efficient Discovery of a Novel CDK20 Small Molecule Inhibitor. *Chem. Sci.* **2023**, DOI: 10.1039/D2SC05709C.
- (29) Ozerov, I. V.; Lezhnina, K. V.; Izumchenko, E.; Artemov, A. V.; Medintsev, S.; Vanhaelen, Q.; Aliper, A.; Vijj, J.; Osipov, A. N.; Labat, I.; West, M. D.; Buzdin, A.; Cantor, C. R.; Nikolsky, Y.; Borisov, N.; Irincheeva, I.; Khokhlovich, E.; Sidransky, D.; Camargo, M. L.; Zhavoronkov, A. In Silico Pathway Activation Network Decomposition Analysis (iPANDA) as a Method for Biomarker Development. *Nat. Commun.* **2016**, *7*, 13427.
- (30) Stamatas, G. N.; Wu, J.; Pappas, A.; Mirmirani, P.; McCormick, T. S.; Cooper, K. D.; Consolo, M.; Schastnaya, J.; Ozerov, I. V.; Aliper, A.; Zhavoronkov, A. An Analysis of Gene Expression Data Involving Examination of Signaling Pathways Activation Reveals New Insights into the Mechanism of Action of Minoxidil Topical Foam in Men with Androgenetic Alopecia. *Cell Cycle* **2017**, *16* (17), 1578–1584.
- (31) Ravi, R.; Noonan, K. A.; Pham, V.; Bedi, R.; Zhavoronkov, A.; Ozerov, I. V.; Makarev, E.; V. Artemov, A.; Wysocki, P. T.; Mehra, R.; Nimmagadda, S.; Marchionni, L.; Sidransky, D.; Borrello, I. M.; Izumchenko, E.; Bedi, A. Bifunctional Immune Checkpoint-Targeted Antibody-Ligand Traps That Simultaneously Disable TGF $\beta$  Enhance the Efficacy of Cancer Immunotherapy. *Nat. Commun.* **2018**, *9* (1), 741.
- (32) Saloura, V.; Izumchenko, E.; Zuo, Z.; Bao, R.; Korzinkin, M.; Ozerov, I.; Zhavoronkov, A.; Sidransky, D.; Bedi, A.; Hoque, M. O.; Koeppen, H.; Keck, M. K.; Khattri, A.; London, N.; Kotlov, N.; Fatima, A.; Vougiouklakis, T.; Nakamura, Y.; Lingen, M.; Agrawal, N.; Savage, P. A.; Kron, S.; Kline, J.; Kowanetz, M.; Seiwert, T. Y. Immune Profiles in Primary Squamous Cell Carcinoma of the Head and Neck. *Oral Oncol.* **2019**, *96*, 77–88.
- (33) Subbannayya, T.; Leal-Rojas, P.; Zhavoronkov, A.; Ozerov, I. V.; Korzinkin, M.; Babu, N.; Radhakrishnan, A.; Chavan, S.; Raja, R.; Pinto, S. M.; Patil, A. H.; Barbhuiya, M. A.; Kumar, P.; Guerrero-Preston, R.; Navani, S.; Tiwari, P. K.; Kumar, R. V.; Prasad, T. S. K.; Roa, J. C.; Pandey, A.; Sidransky, D.; Gowda, H.; Izumchenko, E.; Chatterjee, A. PIM1 Kinase Promotes Gallbladder Cancer Cell Proliferation via Inhibition of Proline-Rich Akt Substrate of 40 kDa (PRAS40). *J. Cell Commun. Signal.* **2019**, *13* (2), 163–177.